Corpus-Based Techniques for Sentiment Lexicon Generation: A Review

Mohammad Darwich, Shahrul Azman Mohd Noah, Nazlia Omar, Nurul Aida Osman
Universiti Kebangsaan Malaysia
Malaysia
{modarwish@hotmail.com} {shahrul@ukm.edu.my} {nazlia@ukm.edu.my}
{p78944@siswa.ukm.edu.my}

JDIM
Journal of Digital Information Management

ABSTRACT: *State-of-the-art sentiment analysis systems rely on a sentiment lexicon, which is the most essential feature that drives their performance. This resource is indispensable for, and greatly contributes to, sentiment analysis tasks. This is evident in the emergence of a large volume of research devoted to the development of automated sentiment lexicon generation algorithms. The task of tagging subjective words with a semantic orientation comprises two core approaches: dictionary-based and corpus-based. The former involves making use of an online dictionary to tag words, while the latter relies on co-occurrence statistics or syntactic patterns embedded in text corpora. The end result is a linguistic resource comprising a priori information about words, across the semantic dimension of sentiment. This paper provides a survey on the most prominent research works that utilize corpus-based techniques for sentiment lexicon generation. We also conduct a comparative analysis on the performance of state-of-the-art algorithms proposed for this task, and shed light on the current progress and challenges in this area.*

Subject Categories and Descriptors:
H.3.1 [Content Analysis and Indexing]: H.2.8 [Database Applications]: Data mining;

General Terms: Sentiment Analysis, Opinion Mining, Sentiment Lexicon, Sentiment Lexicon Generation

Keywords: Sentiment Analysis, Sentiment Lexicon, Sentiment Lexicon Generation, Opinion Mining

Received: 10 May 2019, Revised 5 August 2019, Accepted 14 August 2019

Review Metrics: Review Scale- 0/6, Review Score-5.25, Inter-reviewer Consistency- 88%

DOI: 10.6025/jdim/2019/17/5/296-305

1. Introduction

Sentiment Analysis (SA), or opinion mining (OM), is in essence a natural language processing (NLP) task that involves the detection of user sentiment, attitude, emotion and opinion in natural language text. The (unsupervised) lexicon-based approach involves making use of a sentiment lexicon to compute the global sentiment polarity of a text document, based on the aggregation of the polarity of the individual words embedded within the document (Alqasemi, Abdelwahab, & Abdelkader, 2018; Saif et al., 2017; Fernández-Gavilanes et al., 2016; Hutto, & Gilbert, 2014; Taboada et al., 2011). The primary issue in this approach is that some features are beyond the reach of human introspection, and a supervised classification technique would be able to detect these hidden features. Conversely, the (supervised) classification-based approach involves constructing supervised machine learning classifiers that are fed with manually labeled training data for the classification task (Xing, Pallucchini, & Cambria, 2019). The primary issue with this approach is that it requires manually labeled training data for achieving relatively good accuracy, is

computation-intensive, and naturally possesses a hidden, black-box process. The main task of SA is to classify text units according to their polarity of positive or negative.

SA makes possible a rich set of applications that range from detecting sentiment toward certain topics in the product reviews domain, customer relationship management, the stock market and political figures, among other domains (Chaturvedi et al., 2018; Mantyla et al., 2018; Liu, 2015).

User reviews generated on the Web and social media have become the de-facto standard for measuring the overall quality of products and services (Blair-Goldensohn et al. 2008). Nevertheless, it is a costly and time consuming process for organizations to manually monitor the overwhelmingly massive stream of user generated online product and service reviews. Consequently, organizations often turn to automatic SA models to monitor user sentiment in online reviews, which provides valuable cues for decision making (Liu 2015).

Sentiment words and phrases greatly contribute to, and are an indispensable resource for SA tasks. These are typically compiled in a sentiment lexicon, a linguistic resource comprising a priori information about words, across the semantic dimension of sentiment. In this work, the semantic dimension of sentiment refers to the stereotypical sentiment polarity of the word, or the degree is deviates from the norm, and toward positivity or negativity (Lehrer, 1974). A typical sentiment lexicon has dimensions such as the polarity and strength (or intensity) of the polarity for each word.

However, the problem is that manually tagging words to produce a sentiment lexicon is prohibitively costly in terms of annotator time and effort. Consequently, this area has witnessed the emergence of a large volume of work concentrated on automatic sentiment lexicon generation. The dictionary-based approach involves leveraging lexical resources and online dictionaries (WordNet, Merriam Webster, etc.) to automatically tag terms with their corresponding sentiment polarity (Vicente et al. 2017; Baccianella et al. 2010). Conversely, the corpus-based approach involves exploiting co-occurrence statistics or syntactic patterns in a text corpus (Alqasemi, Abdelwahab, & Abdelkader, 2018; Saif et al., 2017; Deng, Sinha, & Zhao, 2017; Fernández-Gavilanes et al., 2016; Peng, & Park, 2011).Text corpora have been commonly used in domain adaptation, which involves converting a domain-independent sentiment lexicon into a domain-specific lexicon, or domain specific lexicon into an entirely different domain (Alqasemi, Abdelwahab, & Abdelkader, 2018; Saif et al., 2017; Deng, Sinha, & Zhao, 2017; Fernández-Gavilanes et al., 2016). Semi-supervised label propagation has been heavily investigated (Wang et al., 2017; Hamilton et al., 2016; Huang, Niu, & Shi, 2014; Tai, & Kao, 2013; Velikovich et al., 2010). Social media corpora such as Twitter have been utilized to generate informal social media-specific lexicons (Wu et al., 2019; Kimura, & Katsurai, 2017; Tang et al., 2014; Vo, & Zhang, 2016; Severyn, & Moschitti, 2015; Feng et al., 2013; Mohammad, Kiritchenko, & Zhu, 2013; Peng, & Park, 2011). This survey covers an in-depth survey of the mentioned works, as well as other prominent works that use the corpus-based approach for sentiment lexicon generation.

For convenience, the remainder of this paper is structured as follows. Section 2 presents the works that utilize the corpus-based approach to label words in a lexicon with a sentiment polarity. Section 3 discusses the progress made in this area to date, and the challenges that come along with this approach. Section 4 concludes.

2. Corpus-Based Techniques for Sentiment Lexicon Generation

The recent related literature comprises numerous works that have been devoted to the automatic generation of sentiment lexicons using text corpora. The underlying intuition is that the semantic distance between a word and a set of positive and negative seed words can be used as a metric to estimate the sentiment polarity of the target word. This approach relies on co-occurrence statistics or syntactic patterns in text corpora and a set of predefined positive and negative seed words. The information in the context surrounding the target term may be exploited to aid in polarity assignment. This approach is also commonly used to adapt a domain independent sentiment lexicon into a new domain-specific lexicon, using a corpus in the target domain (Alqasemi et al. 2019).

Another important aspect to mention is the dimensions of the sentiment lexicon itself. Typically, a sentiment lexicon comprises a set of words, where each is labeled with a polarity and a strength of the polarity. Other dimensions can include the affective features or emotions underlying the word. This final lexicon is useful and has implications in many domains, such as the product reviews domain.

This new lexicon can then be incorporated into SA models that are designed specifically for sentiment classification tasks on text from that particular target domain. Moreover, unlike the dictionary-based approach, which is confined to the *formal* vocabulary entries in the dictionary or lexical resource used, this approach may pick up *informal* terms and internet slang commonly used in social media text.

The general framework for the construction of sentiment lexicons using text corpora varies, and many techniques have been adopted in the literature. The core corpus-based techniques in the existing literature are discussed hereafter.

2.1 Label Propagation

Velikovich et al. (2010) employ a graph propagation model to semi-automatically derive a polarity lexicon from the Web (as a corpus). The path with the highest weight

between a seed node and a target node was considered to label the target node. Four billion webpages were used to extract 20 million phrases using frequency statistics and mutual information among target terms and predefined seed terms. The advantages of this method are that it is not limited to the coverage of WordNet, and intentionally includes social media text, internet slang and misspellings, which also reflect sentiment. Moreover, it is unsupervised, since only a small set of seeds can be used as input. However, similar to other web-based lexicon generation algorithms (Turney and Littman 2003), it requires a rather large corpus for a suitable recall value.

Huang et al. (2014) generate a domain-specific sentiment lexicon using constrained label propagation. Candidate sentiment words are extracted using POS and chunk dependency parsing. Morphological constraints (*e.g., practical vs impractical*) and pairwise contextual constraints (*e.g., well-appointed and pleasurable*) are extracted from the domain corpus. Domain-independent seed words are extracted from the pros and cons sections of semi-structured reviews. Constraint propagation is performed to spread the effect of local constraints across candidate sentiment words in the corpus.

Tai and Kao (2013) also propose a semi-supervised graph-based label propagation algorithm to generate a domain-specific sentiment lexicon using conjunction rules, SOC-PMI and WordNet. Predefined seed words propagate sentiment scores to unlabeled words. Empirical investigation on a manually labeled test set demonstrates that the method is able to assign or tune word polarity according to the target domain.

Wang et al. (2017) label sentiment terms from within a corpus by applying a concept called neural PU learning, which involves learning from positive unlabeled samples. Wang and Xia (2017) construct a neural network to learn a sentiment-aware word embeddings by applying sentiment supervision at document and term levels, in order to improve the overall quality of a sentiment lexicon. Term level supervision is based on readily available sentiment lexicons or PMI techniques from text corpora. Hamilton et al. (2016) construct domain-sensitive word embeddings with a label propagation approach to generate domain-sensitive sentiment lexicons with small seed sets. They build semantic representations of terms with the corpus with a vector space model.

2.2 Domain Adaptation

Using a corpus from a specific topic or domain helps to tune any lexicon to that particular domain, as demonstrated by prior work that focus on the generation of domain- or context-specific sentiment lexicons. Some prominent work on domain adaptation is highlighted in this subsection.

Labille et al. (2017) proposes an approach for generating a domain-specific lexicon based on a fusion of probabilistic and information theoretic weights. The work is different from the traditional techniques by generating a domain-specific lexicon with no prior knowledge. The effectiveness of several domain-specific lexicons is measured using two gold standard generic lexicons by computing their accuracy. The domain-sensitive lexicons outperform the gold standard lexicons. They demonstrate that text mining techniques perform with comparable accuracy as traditional approaches in the generation of sentiment scores.

Salah et al. (2013) propose two approaches to generating domain-specific sentiment lexicons, namely, direct generation and domain adaptation. The first generates a dedicated lexicon from the labelled source data, while the second uses a general purpose sentiment lexicon and adapts it into a domain-sensitive lexicon on a particular domain. A corpus of labelled political speeches from political debates held within the UK Houses of Commons is used for this task. The primary contributions of the work are the TF-IDF "learning" mechanism used for the labeling of sentiment scores to terms using appropriately defined training data. The second contribution is the technique for performing the desired opinion mining using the generated lexicons.

Fernandez-Gavilanes et al. (2016) employ a fully unsupervised sentiment analysis model that is robust across different domains and contexts. Their approach is based on measuring the dependencies among lemmatized terms with a sentiment propagation algorithm that takes into account various linguistic rules, including negation, intensification, modification and adversative and concessive relationships. Moreover, it is context-sensitive in that sentiment scores are given to terms based on the dependency with neighboring terms that are under the scope of context. They focus on sentiment classification of full-text using this model, but generate sentiment lexicons based on dependency parsing and the context of the term to be labeled. Dependencies between terms were used as edges, and a PageRank algorithm was applied until convergence to label each term with a final sentiment polarity score. A significant benefit of this approach is that the sentiment lexicon generated can be used reliably across multiple domains, since dependency parsing is derived directly 'in real-time' from the actual text content to be classified.

Weichselbraun et al. (2011) use crowd sourcing and a bootstrapping approach to extend sentiment lexicons to a specific domain. Tan and Wu (2011) propose a random walk algorithm using several document- and word-relations from source and target domains. Bollegala et al. (2011) use labeled and unlabeled data from multiple independent domains to determine the association among words that reflect similar sentiment across different domains. Deng et al. (2017) adapt sentiment lexicons for domain-sensitive sentiment classification of social media content, using both a corpus and a dictionary simultaneously. Alqasemi et al. (2018) construct a domain-sensitive sentiment lexicon using KNN search via discrimination vectors. Saif et al. (2017) propose a domain adaptation technique that exploits contextual and semantic information derived from

DBPedia to introduce new terms to a lexicon.

Deng et al. (2018) assign pairs of topics and polarities for individual words in the lexicon. In TaSL, text units are represented by multiple pairs of topics and polarities, and terms are characterized by a multinomial distribution across the pairs of topics and polarities. The primary benefit of TaSL is that the polarities of terms in variable topics can be successfully captured. This model is practical enough to build a topic-sensitive sentiment lexicons. Xing et al. (2019) train a vanilla sentiment classifier model and adapt term polarities to the target domain. They track the incorrectly predicted sentences and apply them as the supervision as an alternative of addressing the gold standard to emulate the life-long cognitive procedure of lexicon learning. An exploration-exploitation technique is constructed to trade-off between updating the polarity of each term, and adding for new subjective terms.

Han et al. (2018) develop a domain-specific sentiment lexicon induction method, whereby mutual information is used to label terms with POS tags within the lexicon, and the training data are chosen from a corpus based on their sentiment scores given by a SentiWordNet classifier.

Mudinas et al. (2018) utilize machine learning algorithms to generate reliable domain-sensitive sentiment lexicons from a handful of predefined sentiment words or seeds. A crucial finding is that simple linear model based supervised learning algorithms can actually work better than more complex transductive learning algorithms that represent modern techniques for sentiment lexicon generation. The lexicon could be used in a lexicon-based approach for polarity classification, but improved performance can be achieved via a two-step bootstrapping method that employed the generated lexicon to label sentiment scores to unlabeled documents in the first step, and then employs those documents to acquire clear sentiment signals as pseudo-labeled examples, in order to train a text polarity classifier with supervised learning algorithms. Generally, the sentiment lexicon needs to be adapted to a particular domain, prior to the supervised classification task, which would in turn contribute to the overall classification task in domain-sensitive scenarios.

Wu et al. (2019) construct a target-specific sentiment lexicon, whereby every term is a sentiment pair consisting of a sentiment target and a sentiment word. An unsupervised algorithm issued to identify sentiment pairs reliably. Both semantic and syntactic features are considered in the algorithm, in order to identify sentiment pairs comprising correct opinion targets. A set of sentiment pairs are generated to classify their polarities. A general-purpose sentiment lexicon, and context knowledge including syntactic relations and sentiment information in sentences, are integrated in a unified model to compute sentiment scores of the sentiment pairs.

Du et al. (2010) and Du and Tan (2009) investigate the problem of domain adaptation from one domain to another; while Choi and Cardie (2009) and Jijkoun et al. (2010) investigate adapting a general-purpose sentiment lexicon into a domain-specific one. Zhang and Liu (2011) develop an approach to assign a sentiment polarity to nouns and noun phrases that belong to a certain target domain.

2.3 Pointwise Mutual Information

Yang et al. (2013) employed a modified variant of the SO-PMI algorithm to generate a sentiment lexicon. For features extraction, the generated sentiment lexicon and the Chi statistic were compared individually, and a naïve Bayes classifier was used to classify a set of Chinese hotel reviews using these two sets of features. They concluded that using PMI co-occurrence statistics performed better compared to other features for the classification task.

Turney and Littman (2003) utilize a small set of paradigm positive and negative seed terms and a bootstrapping algorithm to mark words with a semantic orientation. The intuition is that the orientation of a target word is assigned from the measure of its association (co-occurrence frequency) with a set of known positive words minus the measure of its association with a set of known negative words. The positive and negative seed sets used to define the positive and negative classes are: S_p = {*good, nice, excellent, positive, fortunate, correct, superior*} and S_n = {*bad, nasty, poor, negative, unfortunate, wrong, inferior*} respectively. Note that the seed terms are based on antonymous adjectives (good/bad, nice/nasty, etc.). They first compute the pointwise mutual information (PMI) as follows:

$$PMI\,(term,\, term_i) = log_2 \frac{Pr\,(term,\, term_i)}{Pr\,(term)\,Pr\,(term_i)}$$

where $term$ is the target term and $term_i$ is the seed term. The numerator is the probability both words co-occur together, while the denominator reflects the probability they occur independently. The PMI reflects the measure of the degree of statistical dependence of a word pair, hence, their semantic similarity. The orientation of a target term $O(term)$ is then computed using the seed sets as follows:

$$O(term) = \sum_{term_i \in S_p} PMI\,(term,\, term_i) - \sum_{term_i \in S_n} PMI\,(term,\, term_i)$$

If the value of $O(term)$ is positive, then the term is labeled with a positive orientation, and negative otherwise. The higher the magnitude, or absolute value, of $O(term)$, the stronger the sentiment strength of the term. In the PMI method, term co-occurrence frequencies were computed by sending a query to the AltaVista search engine. A *term* query represents the target term query, a $term_i$ query represents the seed term query, and finally, a $term$ NEAR $term_i$ query represents the query for documents in which they co-occur. The NEAR operator in the search engine returns a document if the terms are located in it at a

proximity of less than ten terms in any order, while the AND operator returns a document if both terms appear together anywhere within the document. Overall, although this approach can be applied to any word class, it is data-intensive, i.e., it requires massive text corpora to yield relatively good accuracy, and is computationally expensive. Moreover, due to varying 'dynamic' search engine results based on added/removed webpages, and also to search results tailored to a particular geographic area, the co-occurrence statistics would vary when repeated. Furthermore, the Altavista search engine no longer supports use of the NEAR operator.

Taboada et al. (2006) also use PMI between a target term and predefined seed terms using the NEAR operator in Altavista and the AND operator in Google. They conclude that Google is not a reliable corpus to be used for word-orientation assignment, and that static 'offline' corpora may be more reliable.

Xu et al. (2012) propose a supervised approach based on Sentiment Hyperspace Analogue to Language (S-HAL) and PMI. Predefined seed terms were used as the base space, and co-occurrence statistics between seeds and target terms defined the orientation of target terms. Based on this model, a binary SVMs classifier was employed to label unseen words and phrases. Evaluation on a manually annotated Chinese test set shows that it outperforms prior PMI methods to measure orientation of words and phrases, without the requirement for search engine queries. Their algorithm forces objective words into the positive and negative categories, and is limited to Chinese.

2.4 Matrix Factorization

Peng and Park (2011) propose a fully automatic method to compile a sentiment dictionary using Constrained Symmetric Nonnegative Matrix Factorization (CSNMF). They leverage a social media corpus in order to take into consideration not only formal words from a dictionary (WordNet), but also informal words (internet slang) commonly used on social media. Most sentiment lexicon generation methods that use the dictionary-based approach to exploit WordNet cover only formal words. In their work, however, they prove that using as resources both a dictionary and a social media corpus in combination allows for the compilation of a more effective lexicon that is able to include both dictionary entries and social media words.

They initially use a seed set to propagate synonym and antonym relations, and update the seed set to 400 words. The final set they obtain from this process ($S_{WordNet}$) is then used as input to the next stage in their approach, which involves two corpora from the social media website *digg.com* (Digg6 and Digg9). They use words extracted from $S_{WordNet}$ to find any adjectives in the corpus that are linked to these words. They apply the conjunctions *and* *but* to find adjectives with equal polarity and opposite polarity, respectively, and label them accordingly (final set now called S).

Two nonnegative symmetric matrices are then used to symbolize *attractions* and *repulsions* among words within *S*. Next, they apply the CSNMF algorithm to cluster words linked by semantic relations to either the positive class or the negative class, and also assign a sentiment strength to each word. Limitations of their corpus-based approach is that numerous computations across a large corpus are required when linking WordNet terms to target terms in the corpus. Moreover, the conjunction rules applied are only limited to adjectives.

2.5 Polar Phrase Extraction

Kanayama and Nasukawa (2006) propose an unsupervised approach to extract polar clauses from a review dataset. They added the concept of intra-sentential and inter-sentential consistency, and assume that neighboring sentences tend to express similar sentiment polarity. The yielded results indicate a 94% accuracy on average. Their approach is robust in varied domains and in the case of a small initial lexicon. Their phrase-level algorithm is limited to Japanese. Wilson et al. (2005) aimed to classify phrases in terms of sentiment polarity and subjectivity using a corpus. Choi and Cardie (2008) and Breck et al. (2007) also dealt with phrase level sentiment classification.

Takamura et al. (2006) propose several models that utilize latent variables for the semantic orientation of two-word phrases (or word pairs). Since they only consider co-occurrence of words in their model, it is language independent.

2.6 Social Media Hashtags and Emoticons

In modern times, people use social media platforms as the primary manner to express their opinions and sentiments towards new products, services, political figures, etc. Therefore, modern sentiment analysis systems must be able to parse free-form social media text. With this, sentiment lexicon generation methods have drastically transformed from using dictionaries, to using social media corpora.

Feng et al. (2013) demonstrate that, in sentiment lexicon generation, using a corpus of Tweets from the microblogging platform Twitter yields higher accuracy compared to three other corpora: Google, Google Web 1T and Wikipedia. They adopt two independent seed sets: *excellent* and *poor*; and 14 paradigm sentiment terms (Turney and Littman 2003). Four independent semantic similarity measures were employed, namely, Dice, Jaccard, Pointwise Mutual Information and Normalized Google Distance.

Two lexicons were used as a gold standard for evaluation: Liu's sentiment lexicon (Liu 2005) and the MPQA subjectivity lexicon (Wilson et al. 2005). For each lexicon (test set), they computed the co-occurrence statistics between each word in the lexicon, and the seed sets, and then applied the four different semantic similarity measures to infer the sentiment polarity of the word. Next, for the Twitter corpus, they repeated this process after

adding two emoticons into the seed set, namely, **:)** and **:(**. The novelty of this work lies in its justification that it is promising to employ a Twitter corpus for labeling words with a polarity, which tends to contain a high volume of subjective content, as compared to Google and Wikipedia, which generally contain mostly factual content.

Mohammad et al. (2013) employ an SVMs classifier to derive sentiment terms from a Twitter corpus. Emoticons and hashtags were used as natural sentiment labels in extracting training data. They yielded an overall F-score of 88.93. Similarly, Pak and Paroubek (2010) categorized Tweets that explicitly contained happy and sad smiley face emoticons, and used them to train a supervised classifier to label terms with a sentiment polarity. Davidov et al. (2010) also use Twitter hashtags and emoticons as training labels for supervised sentiment classification.

Severyn and Moschitti (2015) treat the task of sentiment lexicon construction as a distant supervision problem, and employ an SVMs classifier trained on labeled Tweets to construct a sentiment lexicon. They extract unigrams and bigrams from the Emoticon140 Twitter corpus, and consider a sentiment lexicon, negation, emphatic lengthening, capitalized words, elongated words, multiple punctuation, among others, for features engineering. They claim that adopting off-the-shelf supervised classifiers such as theirs to automatically extract lexicons outperforms other methods such as PMI. However, this is open to debate, since their machine learning model did not take full advantage of the surrounding context of the sentiment terms, and was limited to bigrams.

Wu et al. (2016) propose a Chinese microblog-specific sentiment lexicon generation algorithm. They integrate three types of prior term sentiment knowledge extracted from a 17 million post microblog dataset. First, the PMI between emoticons and unseen terms are computed from a microblog dataset to derive the sentiment score of the unseen terms. Second, sentiment similarity prior knowledge is extracted. Third, prior sentiment knowledge is extracted from readily-available lexicons. To expand coverage, a data-driven approach for new word detection that considers word distribution over text and over users was applied. The assumption was made that if a new word was used by many users, then it would be beneficial to include in the lexicon. Evaluation on two standard Chinese microblog datasets demonstrates its usefulness in both sentiment analysis and subjectivity detection. Their algorithm is limited to Chinese.

Kimura and Katsurai (2017) construct an emoji lexicon by using sentiment terms from the WordNet Affect database, and computing the co-occurrence between the sentiment terms and emojis. Tang et al. (2014) develop a Twitter sentiment lexicon using a representation learning technique. First, a representation learning procedure learns phrase embeddings, which are then represented as features for classification. Next, a seed expansion procedure generates training data for a phrase level sentiment classification model. Vo and Zhang (2016) also construct a Twitter-specific sentiment lexicon using the PMI between terms and social media emoticons to label terms with a polarity.

Bandhakavi et al. (2018) model an emotion corpus for social media sentiment analysis, using a unigram mixture model, combined with an emotion sentiment mapping for the generation of word sentiment lexicons that capture emotion sensitive vocabulary. They evaluate the proposed mixture model in learning emotion sensitive sentiment lexicons with those generated using supervised latent dirichlet allocation (sLDA) as well as word document frequency (WDF) frequency information.

2.7 Conjunction Rules on Adjectives

Hatzivassiloglou and McKeown (1997) were first to focus on the task of automatically assigning adjectives with their corresponding semantic orientation. Their proposed algorithm examines pairs of adjectives that are conjoined by the coordinating conjunctions *and*, *or*, *but*, *either-or*, and *neither-nor*, from the Wall Street Journal corpus. The intuition behind this approach is that conjoining adjectives enforces linguistic constraints on their polarity. The use of *and* tends to conjoin adjectives with equal orientations (e.g., authentic *and* delicious), while the use of *but* conjoins adjectives with opposite orientations (e.g., authentic *but* competitive).

The conjunctions within the training dataset are used to train a log-linear regression classifier that categorizes adjective pairs with either equal or opposing orientations. Their model is limited to labeling adjectives only, since conjunctions do not always enforce linguistic constraints on the sentiment properties between noun pairs (war and peace), or between verb pairs (rise and fall). Moreover, the proposed algorithm is complicated, and requires a considerable amount of training data, and a rather large corpus, to yield a considerable accuracy.

3. Progress and Challenges

Based on the comprehensive review of corpus-based techniques in Section 2, relatively good progress has been made in this area. The corpus-based approach has its advantages over a dictionary-based approach in that it is able to generate a domain-, context-, or topic-sensitive lexicon (Alqasemi et al. 2018), and is able to capture informal terms and internet slang (Peng and Park 2011). However, it does come with limitations, which are mentioned hereafter.

First, unlike a formal dictionary, which readily comprises the entire vocabulary of a natural language, a massive corpus is required in order to capture the entire span of vocabulary words across a natural language. Consequently, although using a corpus is efficient at marking informal slang and social media terms with a polarity, it is inefficient at marking formal terms. Second, a corpus is generally free-form and unstructured, in contrast to the structured

layout of a dictionary. This makes it noisy compared to a formal dictionary.

Third, using a corpus may also be both data- and computation-intensive. For example, Turney and Littman (2003) uses a massive 100 billion word corpus in their PMI algorithm to achieve good accuracy. Fourth, co-occurrence statistics may not always be reliable. For example, Fellbaum et al. (1993; pg. 27) claim that antonyms of adjectives often co-occur together in the same phrases and sentences. Moreover, Kanayama and Nasukawa (2006) mention that only about 60% of co-occurrences reflect similar sentiment. Therefore, using co-occurrence statistics alone as a measure of sentiment polarity is insufficient (Justeson and Katz 1991).

Fourth, another reoccurring issue is with regards to the relationship between sentiment terms and product features in the product reviews domain. Occasionally, a particular sentiment word may be positive towards one product feature, and negative towards another. Therefore, one general solutions is to adapt a domain-independent lexicon based on each feature of a product, prior to classification of the overall sentiment of each of the features of the product found throughout the text.

Fifth, the overall quality of the sentiment lexicon is difficult to measure. In contrast, Schneider et al. (2018) mention that sentiment lexicons that are generated using dictionaries and lexical resources contain complex inaccuracies, beyond the mislabeling of polarity words, which are difficult to manually detect due to the automatic nature of the lexicon generation technique. They also mention that these lexicons exhibit: (a) intra-dictionary inaccuracies, where words are labeled incorrectly; (b) inter-dictionary inconsistencies, where there is contrast between the polarity of words in two different dictionaries; and (c) no consideration of these inconsistencies that occur due to the automatic nature in the approach used to induce them. They attempt to pinpoint inconsistencies found within an individual dictionary, or across multiple dictionaries, with use of a satisfiability problem (SAT). Once these inconsistencies are identified, the lexicon(s) can be improved.

Sixth, an issue regarding existing datasets is that they are not standardized. For example, some data sets contain sentiment rating in formation in the form of five stars, where a one star review reflects a negative review and a five star review reflects a positive review. Other datasets label documents as only positive or negative. Therefore, the dataset used highly depends on the task at hand.

Finally, if a corpus from only one particular domain is available, this would adapt sentiment words to that particular domain, making them unreliable when applied in sentiment classification on a different domain.

4. Conclusion

In this paper we presented a comprehensive review on the notable research works that focus on the corpus-based approach for sentiment lexicon generation. The progress made to date, as well as the challenges inherent in this approach have also been emphasized. The majority of modern sentiment analysis models require domain sensitivity and the consideration of informal cyber text, since social media is now the prominent platform used by people and organization alike to express their opinions and sentiments towards products, political figures, etc. Therefore, corpus-based techniques are now considered a vital part of any modern sentiment analysis system.

References

[1] Alqasemi, F., Abdelwahab, A., Abdelkader, H. (2018). Opinion Lexicon Automatic Construction on Arabic language.

[2] Alqasemi, F., Abdelwahab, A., Abdelkader, H. (2019). Constructing automatic domain-specific sentiment lexicon using KNN search via terms discrimination vectors. *International Journal of Computers and Applications*, 41(2) 129-139.

[3] Andreevskaia, A., Bergler, S. (2008). When specialists and generalists work together: Overcoming domain dependence in sentiment tagging. *In: Proceedings of ACL-08: HLT*, 290-298.

[4] Baccianella, S., Esuli, A., Sebastiani, F. (2010, May). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. *In*: *Lrec*, 10, 2010, 2200-2204.

[5] Bandhakavi, A., Wiratunga, N., Massie, S. (2018). Emotion aware polarity lexicons for Twitter sentiment analysis. *Expert Systems*, e12332.

[6] Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G., Reynar, J. (2008). Building a sentiment summarizer for local service reviews.

[7] Bollegala, D., Weir, D., Carroll, J. (2011, June). Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*- 1, 132-141. Association for Computational Linguistics.

[8] Breck, E., Choi, Y., Cardie, C. (2007, January). Identifying Expressions of Opinion in Context. *In*: *IJCAI*, 7, 2683-2688.

[9] Chaturvedi, I., Cambria, E., Welsch, R. E., Herrera, F. (2018). Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Information Fusion*, *44*, 65-77.

[10] Choi, Y., Cardie, C. (2008, October). Learning with compositional semantics as structural inference for subsentential sentiment analysis. *In*: *Proceedings of the conference on empirical methods in natural language*

processing (p. 793-801). Association for Computational Linguistics.

[11] Choi, Y., Cardie, C. (2009, August). Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. *In*: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing:* 2, 590-598). Association for Computational Linguistics.

[12] Davidov, D., Tsur, O., Rappoport, A. (2010, August). Enhanced sentiment learning using twitter hashtags and smileys. *In*: *Proceedings of the 23rd international conference on computational linguistics: posters* (p. 241-249). Association for Computational Linguistics.

[13] Deng, S., Sinha, A. P., Zhao, H. (2017). Adapting sentiment lexicons to domain-specific social media texts. *Decision Support Systems*, 94, 65-76.

[14] Deng, S., Kwak, D. H., Wu, J., Sinha, A., Zhao, H. (2018). Classifying Investor Sentiment in Microblogs: *A Transfer Learning Approach*.

[15] Du, W., Tan, S. (2009, May). An iterative reinforcement approach for fine-grained opinion mining. *In*: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (p. 486-493). Association for Computational Linguistics.

[16] Du, W., Tan, S., Cheng, X., Yun, X. (2010, February). Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon. *In*: *Proceedings of the third ACM international conference on Web search and data mining* (p. 111-120). ACM.

[17] Esuli, A., Sebastiani, F. (2006, May). Sentiwordnet: A publicly available lexical resource for opinion mining. *In*: *LREC*, 6, 417-422.

[18] Fellbaum, C., Gross, D., Miller, K. (1993). Adjectives in wordnet.

[19] Feng, S., Zhang, L., Li, B., Wang, D., Yu, G., Wong, K. F. (2013). Is Twitter a better corpus for measuring sentiment similarity?. *In*: *Proceedings of the 2013 conference on empirical methods in natural language processing* (p. 897-902).

[20] Fernández-Gavilanes, M., Álvarez-López, T., Juncal-Martínez, J., Costa-Montenegro, E., González-Castaño, F. J. (2016). Unsupervised method for sentiment analysis in online texts. *Expert Systems with Applications*, *58*, 57-75.

[21] Hamilton, W. L., Clark, K., Leskovec, J., Jurafsky, D. (2016, November). Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, vol. 2016, 595. NIH Public Access.

[22] Han, H., Zhang, J., Yang, J., Shen, Y., Zhang, Y. (2018). Generate domain-specific sentiment lexicon for review sentiment analysis. *Multimedia Tools and Applications*, *77*(16) 21265-21280.

[23] Hatzivassiloglou, V., McKeown, K. R. (1997, July). Predicting the semantic orientation of adjectives. *In*: *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the European chapter of the association for computational linguistics* (p. 174-181). Association for Computational Linguistics.

[24] Hu, M., Liu, B. (2004, August). Mining and summarizing customer reviews. *In*: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (p. 168-177). ACM.

[25] Huang, S., Niu, Z., Shi, C. (2014). Automatic construction of domain-specific sentiment lexicon based on constrained label propagation. *Knowledge-Based Systems*, 56, 191-200.

[26] Hutto, C. J., Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *In*: *Eighth international AAAI conference on weblogs and social media*.

[27] Jijkoun, V., de Rijke, M., Weerkamp, W. (2010, July). Generating focused topic-specific sentiment lexicons. *In*: *Proceedings of the 48th annual meeting of the association for computational linguistics* (p. 585-594). Association for Computational Linguistics.

[28] Justeson, J. S., Katz, S. M. (1991). Co-occurrences of antonymous adjectives and their contexts. *Computational linguistics*, 17(1) 1-19.

[29] Kanayama, H., Nasukawa, T. (2006, July). Fully automatic lexicon expansion for domain-oriented sentiment analysis. *In*: *Proceedings of the 2006 conference on empirical methods in natural language processing* (p. 355-363). Association for Computational Linguistics.

[30] Kimura, M., Katsurai, M. (2017, July). Automatic construction of an emoji sentiment lexicon. *In*: *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017* (p. 1033-1036). ACM.

[31] Labille, K., Gauch, S., Alfarhood, S. (2017, August). Creating domain-specific sentiment lexicons via text mining. *In*: Proc. Workshop Issues Sentiment Discovery Opinion Mining (WISDOM).

[32] Lehrer, A. (1974). Semantic fields and lexical structure.

[33] Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

[34] Mäntylä, M. V., Graziotin, D., Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review*, 27, 16-32.

[35] Mohammad, S. M., Kiritchenko, S., Zhu, X. (2013). NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.

[36] Mudinas, A., Zhang, D., Levene, M. (2018). Bootstrap domain-specific sentiment classifiers from unlabeled corpora. *Transactions of the Association of Computational Linguistics*, 6, 269-285.

[37] Pak, A., Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. *In*: *LREc*, 10 (2010) 1320-1326.

[38] Pang, B., Lee, L., Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. *In*: *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-* 10, 79-86. Association for Computational Linguistics.

[39] Pang, B., Lee, L. (2005, June). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *In*: *Proceedings of the 43rd annual meeting on association for computational linguistics* (p. 115-124). Association for Computational Linguistics.

[40] Peng, W., Park, D. H. (2011, July). Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization. *In*: *Fifth International AAAI Conference on Weblogs and Social Media*.

[41] Saif, H., Fernandez, M., Kastler, L., Alani, H. (2017). Sentiment lexicon adaptation with context and semantics for the social web. *Semantic Web*, 8(5) 643-665.

[42] Salah, Z., Coenen, F., Grossi, D. (2013, December). Generating domain-specific sentiment lexicons for opinion mining. *In*: International Conference on Advanced Data Mining and Applications (p. 13-24). Springer, Berlin, Heidelberg.

[43] Schneider, A., Male, J., Bhogadhi, S., Dragut, E. (2018). DebugSL: An Interactive Tool for Debugging Sentiment Lexicons. *In*: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations.
36–40.

[44] Severyn, A., Moschitti, A. (2015). Unitn: Training deep convolutional neural network for twitter sentiment classification. *In*: *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)* (p. 464-469).

[45] Taboada, M., Anthony, C., Voll, K. D. (2006, May). Methods for Creating Semantic Orientation Dictionaries. *In*: LREC (p. 427-432).

[46] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, *37*(2) 267-307.

[47] Tai, Y. J., Kao, H. Y. (2013, December). Automatic domain-specific sentiment lexicon generation with label propagation. *In*: *Proceedings of International Conference on Information Integration and Web-based Applications & Services* (p. 53). ACM.

[48] Takamura, H., Inui, T., Okumura, M. (2006). Latent variable models for semantic orientations of phrases. *In*: *11th Conference of the European Chapter of the Association for Computational Linguistics*.

[49] Tan, S., Wu, Q. (2011). A random walk algorithm for automatic construction of domain-oriented sentiment lexicon. *Expert Systems with Applications*, 38(10) 12094-12100.

[50] Tang, D., Wei, F., Qin, B., Liu, T., Zhou, M. (2014). Coooolll: A deep learning system for twitter sentiment classification. *In*: *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)* (p. 208-212).

[51] Turney, P. D., Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4) 315-346.

[52] Velikovich, L., Blair-Goldensohn, S., Hannan, K., McDonald, R. (2010, June). The viability of web-derived polarity lexicons. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (p. 777-785). Association for Computational Linguistics.

[53] Vicente, I. S., Agerri, R., Rigau, G. (2017). Q-wordnetppv: Simple, robust and (almost) unsupervised generation of polarity lexicons for multiple languages. *arXiv preprint arXiv:1702.01711*.

[54] Vo, D. T., Zhang, Y. (2016). Don't Count, Predict! An Automatic Approach to Learning Sentiment Lexicons for Short Text. *In*: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2, 219-224.

[55] Wang, W., Pan, S. J., Dahlmeier, D., Xiao, X. (2017, February). Coupled multi-layer attentions for co-extraction of aspect and opinion terms. *In*: *Thirty-First AAAI Conference on Artificial Intelligence*.

[56] Wang, L., Xia, R. (2017, September). Sentiment lexicon construction with representation learning based on hierarchical sentiment supervision. *In*: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (p. 502-510).

[57] Weichselbraun, A., Gindl, S., Scharl, A. (2011, October). Using games with a purpose and bootstrapping to create domain-specific sentiment lexicons. *In*: *Proceedings of the 20th ACM international conference on Information and knowledge management* (p. 1053-1060). ACM.

[58] Wilson, T., Wiebe, J., Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *In*: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

[59] Wu, F., Huang, Y., Song, Y., Liu, S. (2016). Towards building a high-quality microblog-specific Chinese sentiment lexicon. *Decision Support Systems*, 87, 39-49.

[60] Wu, S., Wu, F., Chang, Y., Wu, C., Huang, Y. (2019). Automatic construction of target-specific sentiment lexicon. *Expert Systems with Applications*, 116, 285-298.

[61] Xing, F. Z., Pallucchini, F., Cambria, E. (2019). Cognitive-inspired domain adaptation of sentiment lexicons. *Information Processing & Management*, 56(3) 554-564.

[62] Xu, T., Peng, Q., Cheng, Y. (2012). Identifying the semantic orientation of terms using S-HAL for sentiment analysis. *Knowledge-Based Systems*, 35, 279-289.

[63] Yang, A. M., Lin, J. H., Zhou, Y. M., Chen, J. (2013). Research on building a Chinese sentiment lexicon based on SO-PMI. *In*: *Applied Mechanics and Materials*, 263, 1688-1693. Trans Tech Publications.

[64] Zhang, L., Liu, B. (2011, June). Identifying noun product features that imply opinions. *In*: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers*- 2, 575-580. Association for Computational Linguistics.